An X-ray database, tools and procedures for the study of speech production

Rudolph Sock, Fabrice Hirsch, Yves Laprie, Pascal Perrier, Béatrice Vaxelaire, Gilbert Brock, Fayssal Bouarourou, Camille Fauth, Véronique Ferbach-Hecker, Liang Ma, et al.

[1]DOCVACIM

An X-ray database, tools and procedures for the study of speech production

Rudolph Sock[1], Fabrice Hirsch[2], Yves Laprie[3], Pascal Perrier[4], BéatriceVaxelaire[1], Gilbert Brock[1], Fayssal Bouarourou[1], Camille Fauth[1], VéroniqueFerbach-Hecker[1], Liang Ma[4], Julie Busset[3] and Jean Sturm[1]

[1]Université de Strasbourg, Institut de Phonétique de Strasbourg – IPS & U.R. 1339 Linguistique, Langues et Parole – LilPa, E.R. Parole et Cognition

[2]Université Paul Valéry - Montpellier III, Praxiling UMR 5267, CNRS

[3]LORIA/CNRS UMR 7503 Villers-les-Nancy, France

[4]Gipsa-Lab de Grenoble, Département Parole et Cognition, UMR CNRS 5009

sock@unistra.fr

Abstract. *This paper presents an X-ray database and processing tools that have been respectively elaborated and developed within a research project on speech production (DOCVACIM). The X-ray data deal with various phonetic issues in different languages spoken in Europe, in Africa, in Asia and in Latin America. The goal of the project, apart from looking into issues related to coarticulation, inversion and evalustion of physical models, is to make available to the speech scientific community: 1) a set of multilingual and multimedia data on speech production containing cine-radiographic images of the vocal tract, acoustic signals, tracings and contours of the vocal tract, all synchronised and accessible within a processing platform; 2) adapted tools and software that allow extracting articulatory information from these data, in relation with prior phonological labelling. The focus in this article is on the adapted tools and software which have been developed within the project.*

1. Introduction

Today, Electro Magnetic Articulometers (EMA), optical motion capture systems, ultrasounds and dynamic MRI are the favoured experimental setups used to study speech kinematics. EMA and optical systems have the strong advantage in that they

[1] DOCVACIM : DOnnées Cinéradiographiques Valorisées pour l'étude de la Coarticulation, de l'Inversion et l'évaluation de Modèles physiques. An Agence Nationale de la Recherche (ANR n° ANR-07-CORP-018) Project, 2007 - 2011

perform at sample frequencies which are way above the Nyquist frequency of vocal tract movements. They also have the crucial possibility to inform on actual flesh-point movements. However, optical systems are obviously not adapted to visualisation of non-visible movements that occur inside the vocal tract. Similarly, EMA systems provide information limited to the region extending from the lips to the rear part of the hard palate. To get information on the remaining part of the vocal tract, ultrasounds and MRI techniques are more appropriate, in spite of the fact that they offer a global description without any information on flesh points. However, ultrasound data are also spatially limited by the shadow projected by the hyoid bone onto the lower part of the pharynx, and most current systems provide data at the limited sampling rate of 30Hz. In addition, calibration issues are still under debate regarding such systems. MRI techniques provide very accurate views of vocal tract geometry. Nevertheless, even if constant improvements suggest that such techniques will be the best in a near future, 2D dynamic MRI cannot, to day, provide images at a rate higher that 20 Hz. In addition, the latter technique calls for an expensive experimental setup, which will not necessarily be available to every speech laboratory. In this context, endeavouring for the dissemination of cine-radiographic data recorded in the past, at a time when no regulations limited their use, and of tools allowing their analyses and their application in the context of speech motor control studies, and of articulatory modelling approaches, represents an interesting contribution to problems related to the speech sciences in general. Cine-radiographic systems provide 2D images of the whole vocal tract, from the glottis to the lips at 50 or 60Hz rates.

The Phonetics Institute of Strasbourg (IPS), whose research activities, in part, are conducted within a larger laboratory, LiLPa (Linguistique, Langues et Parole), has been the centre of excellence in France, and one of the main centres in the world, for cine-radiographic data acquisition. This is mainly due to the expertise of Georges Straka, Pela Simon (Simon, 1967) and André Bothorel (Bothorel *et al.*, 1986). In this context, the current project, coordinated by the IPS in Strasbourg, the Gipsa-Lab in Grenoble and the LORIA in Nancy, aims to share some of the best X-ray movies on speech production that were made at the IPS, as from the end of the fifties. The project concerns 20 movies of high quality, dealing with linguistic issues in languages spoken in Europe, in Africa, in Asia and in Latin America. The following will be made available to the scientific community: 1) a set of multilingual and multimedia data, unique in the world, on speech production containing cine-radiographic images of the vocal tract, acoustic signals, tracings and contours of the vocal tract, all synchronised and accessible within a processing platform; 2) adapted tools and software that allow extracting articulatory information from these data, in relation with prior phonological labelling.

2. Tools and procedures developed

Despite their intrinsic interest, X-ray films cannot be used directly. We therefore developed software intended to extract, store and process contours of speech articulators. This software called “X-articulators” comprises tools to, firstly, delineate contours and, secondly, to exploit them.

2.1. Delineating articulator contours

Drawing contours by hand is a tedious task and several works have been dedicated to the automatic tracking of articulator contours (Thimm and Luettin, 1999; Laprie and Berger, 1996; Jallon and Berthommier, 2009). We developed software, called "X-articulators"', enabling several tracking tools to be used according to the nature of articulators.

Rigid structures, like the mandible, can be tracked robustly by correlating a reference image, with images to process. In this particular case, the photometric energy of the mandible is sufficiently stronger than that of the tongue to neglect its influence all the more since the region to be correlated can be chosen so as to minimize the effect of the tongue. This simple tracking turned out to be very efficient to tract the mandible, the upper part of the skull in order to compensate for head movements, and even the hyoid bone provided that it does not move too high and intersect other high photometric organs. Tracking provides the displacement parameters, *i.e*. rotation and translation.

Structure deforming along time, and giving rise to one contour only, *i.e.* the lips, the larynx and the epiglottis, have been tracked via the algorithm proposed by Jallon and Berthommier (2009). Contours have been drawn by hand for a series of key images (approximately 10% of the total number of images to process) and approximated by B-spline curves, with a constant number of control points.

The general idea is to choose a rectangular region in the image including the object to be tracked. Then, images are cut to keep only this region in order to remove the influence of other organs. Non key images are indexed by calculating the distance of their DCT coefficients (Discrete Cosine Transform) to those of the key images. For each image, the three closest key images are kept with their distance to the image analyzed. Finally, the new contour given by the control points of the spline is the weighted average of the contours of the three closest key images. If the visual evaluation of tracking shows that some images are not tracked correctly, because they are too far from key images, then they are added as key images.

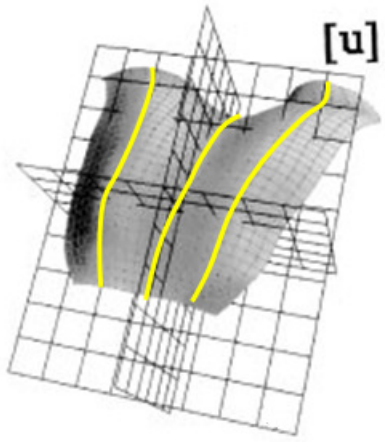

Figure 1: Tongue (from Lundberg and Stone (1999) with the three contours (in yellow) that can appear on an X-ray image.

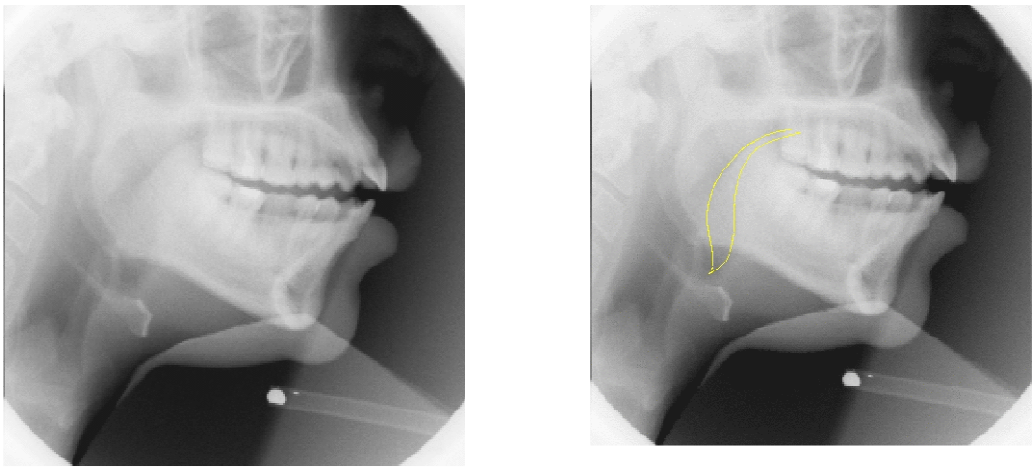

Figure 2: Examples of contours on X-ray images. The two yellow lines in the right image correspond to the tongue contours visible in the pharyngeal cavity.

The tongue contour is much more difficult to track automatically since there are one or two contours depending on tongue shape. The contour to consider is that of the mid-sagittal plane (see Vaxelaire et al., 2009 for a discussion). In many images, the tongue presents a marked groove located at the mid-sagittal plane, which gives the relevant contour. The two upper edges of the tongue may give rise to a second image contour (or exceptionally two if the tongue is not symmetric at all) (see Fig. 1). In practice, this explains the existence of two contours in X-ray images. Since we also recorded MRI images for one speaker of the database it was possible to check contours chosen to represent the tongue even if the tongue was partially hidden. Another difficulty is the presence of teeth in the oral cavity which hide the tongue contour. Human experts often use a completion strategy consisting in drawing a convex contour in the mouth probably

because this shape is much more natural. However, ultrasound images, and MR images as well, show that the tongue contour is concave not only for retroflex tongue shapes but also for many other less extreme tongue shapes.

Although there are some automatic or semi-automatic tracking algorithms, the difficulties presented above led us to draw tongue contours by hand in order to guarantee their relevance. However, the graphical interface of X-articulators offers many tools to make this work easier. In particular, it is well known that the perception of movement by eyes makes the detection of contours easier. We thus added tools to play a small number of contiguous images, preceding or following the current image, when drawing a contour.

2.2. Processing of articulator contours

Contours alone are not very useful and therefore X-articulators provide several tools to make the exploitation of contours easier.

Firstly, it is possible to add landmarks on contours in order to identify a given anatomical point (tongue apex, higher incisor…) used to anchor a coordinate system (e.g. that of the articulatory model). Landmarks can also be used to link two objects. For instance the specific region of the mandible used to track this structure has been chosen to minimize interactions with the tongue and aluminium filters. Its shape thus does not correspond to that of the mandible. A more realistic contour of the mandible can be attached to this region and its movement is copied from the region used for tracking. Secondly, since it is important to relate contours to phonemes uttered, it is possible to import a file of phonetic annotations. Synchronization of the annotation is carried out via visual events which produce an acoustic event simultaneously.

Thirdly, X-articulators provides tools to construct articulatory models from the articulator contours via Principal Component Analysis (PCA). The main articulator is the mandible, whose movement is approximated by two linear components controlling the rotation and the shift. The first component roughly corresponds to jaw aperture. The deformation modes of the tongue are then determined by PCA after the mandible movement has been subtracted from the tongue contours. Then, lip deformations are approximated by two factors (roughly aperture and protrusion), and the lower part of the pharynx (including the larynx and the epiglottis) is represented by two factors. The entire model is described in Laprie and Busset (2011).

2.3. Semi-automatic articulatory measurement

A semi-automatic procedure to measure displacements of articulators of the vocal tract from radiographic tracings has been defined, implemented and validated. For such an analysis, it is important to extract a limited number of points which inform, in a relevant manner, on anterior-posterior and vertical positions of articulators, and from which, in the absence of flesh-point information, and by calculating velocities and accelerations, would adequately reflect the kinematic behaviour of vocal tract articulators. In this aim, we have proposed to describe the tongue contour from, on one hand, the three most elevated intersection points of the superior contour of the tongue, using 4 horizontal

lines, parallel to the average direction from the hard palate, and spaced every 0.5 cm (the highest line going through the superior incisive), and, on the other hand, the 3 most posterior intersection points of the lingual contour using 4 lines parallel to the average direction from the posterior pharyngeal wall (the most retreated being that which approaches the pharyngeal wall). This method is currently being tested in the kinematic analysis of the effects of juncture in one of the French movies.

3. The database and accessing the database

The database on the site currently comprises 4 X-ray-movies, chosen from the archives of the IPS; they have been digitized and synchronized with the audio. Access to the database is at present restricted, as it is in an assessment phase, carried out by laboratories that have expertise in the study of speech production, based on X-ray data. Consulting and distribution of data to a larger public could only be done after refinement of the ergonomics of the database and clarification of some legal issues related to distribution of processed medical data.

Procedures to access the data, via internet, have also been defined. A restricted sub-set can be visualized directly at the website of the project.

4. Searching in the database

Several searching methods are available (Figure 3):

1. Search by *corpus*: the user types the name of the researcher for whom the movie was made, and in response all sentences of the entire corpus will be listed;
2. Search by *sequence*: the user types a string of sounds, and the database in return gives all sentences which comprise this string of sounds, side-by-side;
3. Search by *phoneme*: the user selects a sound or sounds he or she wants to examine. The base proposes all sentences in which the sound or the sounds requested figure. In case of several sounds, the base will present all sentences in which these sounds appear, be they side-by-side or not.
4. Search by *language*: the user selects the language he or she wants to study. The base proposes all sentences in the selected language.
5. A cross-research, using all these different criteria is also possible.

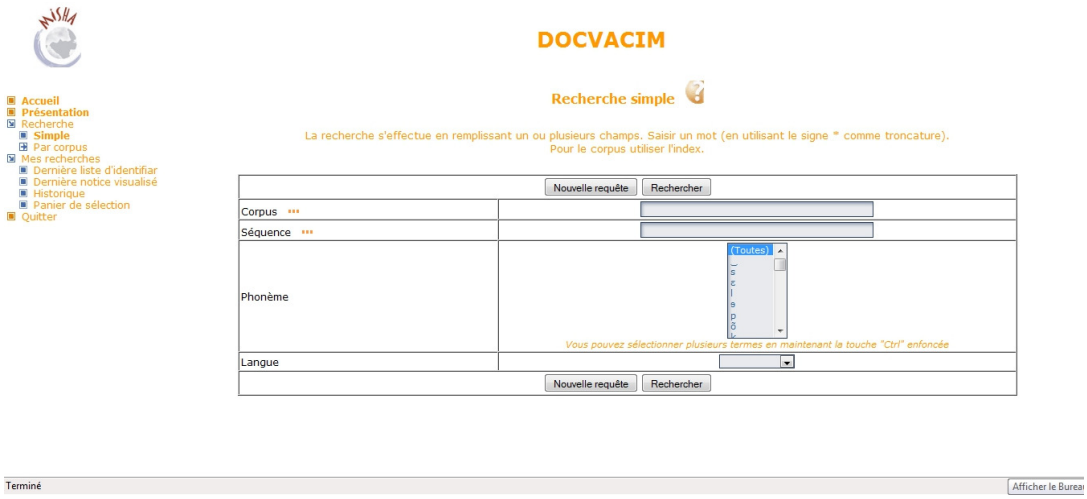

Figure 3. The database form. Searches can be made by corpus, sequence, phoneme and language.

5. Sheets and selecting a sheet

A sheet is available and searchable for each sentence in the database (Figure 4). It furnishes information on the speaker, the language and the different phonemes which appear in the sentence. Some images that allow the user to verify the quality of the images of the movie are visible. When tracings of the vocal tract are available, they are also available in the sheet. Finally, note that the movie could be played.

The user may select a sheet or several sheets, and place them in his or her basket, so as to be able to retrieve later and more easily movies of interest.

Figure 4. Example of a sheet. Informations on the speaker, the language and different phonemes which appear in the selected sentence are given. The movie can be played to evaluate its quality; tracings of the vocal tract area are proposed here.

6. Ordering a movie

Movies on the site are of poor quality. To obtain them in a better resolution, an e-mail message should be sent to sock@unistra.fr. The movies would then be sent by internet or on a DVD (in a *DV*, *.avi* or *.mov* format).

7. Conclusion

The IPS X-ray data correspond to 20 movies of high quality. Some of these movies are associated with digitised hand-drawn tracings. Others are currently being processed with the help of the software X-articulators.

By the time of the ISSP conference, 10 movies, together with available associated data would be sent, on request, to researchers preoccupied with speech production and perception studies.

References

Bothorel, A. Simon, P. Wioland, F. Zerling, J.-P. Cinéradiographie des voyelles et des consonnes du français. Recueil de documents synchronisés pour quatre sujets : vues latérales du conduit vocal, vues frontales de l'orifice labial, données acoustiques. Travaux de l'Institut de Phonétique de Strasbourg, 296 P., 1986.

Laprie, Y. and Berger, M. Towards automatic extraction of tongue contours in x-ray images. In *Proceedings of International Conference on Spoken Language Processing*, volume 1, pages 268–271, Philadelphia (USA), October 1996.

Thimm, G. and Luettin, J. Extraction of articulators in x-ray image sequences. In *Proc.EUROSPEECH*, pages 157–160, Budapest, September 1999.

Lundberg A.J. and Stone M. Three-dimensional tongue surface reconstruction: practical considerations for ultrasound data. Journal of the Acoustical Society of America, 106(5):2858-2867, 1999.

Jallon, J. F. and Berthommier, F. A semi-automatic method for extracting vocal-tract movements from x-ray films. *Speech Communication*, 51(2):97–115, 2009.

Laprie Y. and Busset J.In Proceeding of International Seminar on Speech Production, Montreal, June 2011.

Simon, P. Les consonnes françaises. Mouvements et positions articulatoires à la lumière de la radiocinématographie. Paris, Klincksieck, 380 p., 1967.

Vaxelaire, B. Marchal, A. Hirsch, F. Sock, R. Apports des techniques radiologiques et de la radiocinématographie à l'étude de la production de la parole. In A. Marchal and C. Cavé (eds.), Imagerie médicale pour l'étude de la parole, Hermès Sciences, 125-145, 2009.